What is a Grid?

Grid Today, AUGUST 12, 2002: VOL. 1 NO. 9

(http://www.gridtoday.com/02/0812/020812.html)

I would like to provide perspective on the question of what is a Grid - a perspective derived from several years of building production Grids.

For a significant segment of the Grid community, most of whom are represented at the Global Grid Forum, the primary significance of Grids is that they will provide a persistent and uniform infrastructure for solving complex science and eCommerce problems based on a dynamic and distributed set of resources.

A Grid is an environment that provides access and management for the whole range of computing resources needed to solve complex computing and data handling problems. A Grid is not batch schedulers, cluster managers, or storage systems that happen to be connected to the Internet. Rather, a Grid is a well understood and standardized set of services that provide uniform access to a large number of diverse and distributed resources, together with several critical auxiliary services for resource discovery and secure communication based on authenticated, global identity.

There are a collection of people from major scientific collaborations and computing centers and that have been working for the past two to four years to build production computing and data management Grids to support large-scale scientific collaborations: That is, scientific endeavors that involve world-wide collections of people that have a common goal, and must use shared computing and data resources to achieving that goal.

From this substantial effort to build a common and persistent Grid infrastructure from a dynamic set of underlying resources, an understanding is emerging as to the minimal set of functions needed support large-scale production Grids. (By “production Grids” we mean the Grids that must support a diverse user community to whom the operators of the Grids are responsible for providing a persistent, reliable, and useful service.)

Defining and understanding these minimal services is very important for several reasons. First they represent the fundamental persistent infrastructure and services of Grids that application developers can rely on. Second, they assure that applications can move and/or interoperate among Grids (because the basic services will be available everywhere). Third, they represent the aspects of Grids that require almost all of the operational support needed to provide a useful infrastructure. (This operational support represents most of the cost of providing Grids due to the training and day-to-day attention by the operational staff of the computing and data handling facilities needed to support these services.)

Based on our work in building production Grids, our current understanding of the minimal Grid Functions includes the following set of services that knit together distributed, heterogeneous components:

- resource discovery (services that locate the resources that are suitable for solving a particular problem)
- resource scheduling (coordination of multiple resources so that they may work cooperatively on the same task)
- uniform computing access
- uniform data access
- asynchronous information sources (events, monitoring, logging, and auditing)
- authentication, delegation, and secure communication (the basic Grid security services)
- identity certificate management (the basis of trust in large-scale, widely distributed collaborations)
- system management and access

All of these functions have elements that must be installed and managed on the Grid resources / systems, or have independent servers that provide the functions (e.g. discovery and identity certificate management).

This minimal set of functions represents the services that provide the resource independence and access that will make the Grid a common infrastructure for all higher-level services. That is, they are the smallest set of services that are needed to build all other Grid frameworks, middleware, and applications. (The minimal services may vary somewhat depending on the type of Grid resource. That is, computing resources (super computer, cluster, individual systems), data resources (any managed dataset, database, tertiary storage system, etc.), instrument resources (sources of data based on observing or manipulating the real world), etc., may have a slightly different set of these services, but these are just subsets of the same set of "minimal" functions. For example, you probably cannot initiate computation on most Grid based instrument control systems. However, if you could, it would through the same Grid function as any other computational resource.)

The issue of uniformity deserves a few more comments. We have to be able to move, and / or interoperate, applications, frameworks, and middleware from one Grid to another. We have currently demonstrated good interoperability between such large-scale Grids as NASA's IPG, the NCSA and SDSC NSF PACI Grids, the DOE Science Grid, etc. (whose combined scientific computing capacity represents a significant fraction of the US total). However, this interoperability has come from the common use of the Globus toolkit. One of the main functions of the GGF is to define and standardize protocols for the functions mentioned above so that anyone can write a compatible Grid component based on these definitions, and we can achieve interoperability beyond Globus.

The scope of these minimal Grid services, and their architectural relationship to other Grid services, is being worked out in the Global Grid Forum, Grid Protocol

Architecture Working Group, whose notes and documents may be found at http://www.itg.lbl.gov/GPA/

William Johnston, Lawrence Berkeley National Lab and NASA Ames Research Center (wejohnston@lbl.gov)

Co-Director (with Jeff Nick, IBM), GGF Architecture Area, and co-chair (with Ian Foster and Reagan Moore) Grid Protocol Architecture Working Group.